DOCUMENT RESUME

ED 251 804 CS 007 860

AUTHOR Lyman, Barbara Gallow
TITLE The Construction of a Criterion-Referenced Test of Reading for College Freshmen: An Overview of Theoretical Considerations.
PUB DATE [83]
NOTE 16p.
PUB TYPE Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; *Diagnostic Tests; *Educational Assessment; Higher Education; *Reading Diagnosis; *Reading Tests; Student Placement; *Test Construction; Testing Problems; Test Theory

ABSTRACT

Questions about the appropriateness of the Nelson-Denny Reading Test have led to consideration of the use of a criterion referenced test for assessing students' reading abilities before and after each semester of developmental reading instruction. In constructing such a test, a number of steps should be followed. The first step is to clarify the instructional intent, including such particulars as what the student is doing when demonstrating achievement, what the conditions are under which the student is to perform, and how well the student is to perform. The next is to specify domain, during which those who favor objectives-referencing must write either a single behavioral objective or a set of learning objectives specifying the precise skill(s) and mastery level(s) expected of each student. Next, test specifications need to be written as well as test items, and test items need to be edited. Following this, test items should be reviewed for content validity, to insure congruence or consistency between items and the objectives from which they are drawn. Based on this data, the test items are once again edited and the test assembled. The final steps involve setting performance standards, pilot testing, assessing validity and reliability, and preparing a user's manual. Beyond this final step there is a need to continue to gather technical data as the test is used in different situations and with different student populations. (HOD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☐ This document has been raproduced as received from the person or organization originating it.
☒ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED251804

The Construction of a Criterion-Referenced Test of Reading for College Freshmen: An Overview of Theoretical Considerations

Barbara Gallow Lyman

University of Southwestern Louisiana

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Barbara G. Lyman
TO THE EDUCATIONAL RESOURCES

The Construction of a Criterion-Referenced Test of Reading for College Freshmen: An Overview of Theoretical Considerations

Despite the undeniable advantages of norm-referenced tests for measuring the reading proficiency of entering college students, developmental reading educators question aspects of these tests and of the use to which their results are put. These concerns have generated efforts to create alternatives. At the University of Southwestern Louisiana, for instance, the Nelson-Denny Reading Test has been in use for over a decade. The test is employed as a placement instrument administered at the beginning of each semester to aid in determining which entering freshmen are required to enroll in a developmental reading course. The Nelson-Denny is also used as one means of assessing reading growth which takes place among students as a result of enrollment in developmental reading courses. While long use of this test has revealed its practicality for the purposes designated above, some limitations of the test have also become apparent. Questions about its suitability as a diagnostic instrument have led to a project to develop a criterion-referenced test to supplement, if not to replace, the Nelson-Denny.

Southwestern's developmental reading program is not alone in relying on the Nelson-Denny Reading Test. It enjoys wide use among developmental reading programs at public colleges and universities within the state of Louisiana. According to a 1982 survey conducted by the developmental reading faculty at Louisiana State University,

12 of 18 state colleges, universities, and their various branches were using this instrument to determine students' entry into and exit from their developmental reading programs. Programs which did not rely on the Nelson-Denny for making course placement and exit decisions were nevertheless using it to measure student progress.

The seeming ubiquity of the Nelson-Denny in college developmental reading programs in the state attests to its utility for examining large groups within limited periods. This wide use of the test does not, however, signify that it is the ideal instrument for discovering the strengths and weaknesses of developmental reading students. For example, a number of limitations were cited by Cummins in a 1981 Journal of Reading review of the then newly published Forms E and F of the Nelson-Denny (Brown, Bennett & Hanna, 1981). Although Cummins confined his discussion to difficulties with the Nelson-Denny, the problems he observed represent those encountered in other norm-referenced multiple-choice tests of reading. Concerning the Vocabulary subtest of the Nelson-Denny, Cummins found the contextual cueing at times not well executed, "leaving an item open to two arguably acceptable answers" (p. 55). Another observation about individual test items was that "there is also the occasional item that requires a student to make preposterously minute discriminations between two answer choices that appear to be equally correct" (p. 55). Mathematician and test critic Banesh Hoffman would not be surprised by such charges. In his book, The Tyranny of Testing, Hoffman (1962) maintains that ambiguity is an inherent quality of tests constructed in multiple-choice format. Hoffman further asserts that the most capable students are at the

greatest disadvantage when faced with these instruments because it is they who are most likely to perceive and to be troubled by ambiguity among response choices. Hoffman's greatest objection is that multiple-choice tests do not allow subjects to give reasons for their answers. The idea that norm-referenced multiple-choice tests are flawed instruments is echoed by writer David Owen (1983) in his analysis. While he concentrates on college entrance testing, Owen addresses common concerns about norm-referenced multiple-choice testing such as those elaborated by Cummins in his review of the 1981 forms of the Nelson-Denny.

In addition to difficulties noted regarding less than perfect test items, Cummins pointed out that because of the test's limited time format, "low Vocabulary and Comprehension scores may not always reflect a subject's low achievement in these areas but may be the result of slow cognitive processing, perceptual deficits, or test anxiety" (p. 57).

Aside from the problems of test-item ambiguity and the seeming inability of the Nelson-Denny to distinguish between the slow cognitive processor and the genuinely weak reader, another aspect of the test which limits its usefulness has to do with its being essentially a global test of proficiency. As such the instrument does not offer a great deal of insight into the specific strengths and weaknesses of subjects. Item analyses which can be accomplished based on information provided to examiners in the manual result in quite general, if not superficial, diagnoses. It is not particularly useful for either diagnostic or instructional purposes to know about the 100 Vocabulary subtest items that 50 are prefixed and 50 non-prefixed. While there may be greater value in the distinction made among Comprehension subtest items identifying 18 of the 36 as literal items and the remaining

18 as interpretive comprehension items, these labels also yield information of limited range and depth.

Although the Nelson-Denny has served Southwestern's developmental reading program well by virtue of being a very practical instrument to administer and score, questions about its appropriateness have led to an investigation of other means of assessing students' reading abilities before and after each semester of developmental reading instruction. The alternative to norm-referenced testing which is being explored is criterion-referenced testing. The remainder of this paper sets forth a summary of research into theoretical considerations in criterion-referenced test construction.

In the initial stages of researching the literature, it became apparent that it was not only logical but also necessary to sift through various definitions for the concept of criterion-referenced testing.

The term was first used by Glaser (1963) in his seminal essay on the subject in which he wrote:

> The scores obtained from an achievement test provide primarily two kinds of information. One is the degree to which the student has attained criterion performance, for example, whether he can satisfactorily prepare an experimental report, or solve certain kinds of word problems in arithmetic. The second type of information that an achievement test score provides is the relative ordering of individuals with respect to their test performance, for example, whether Student A can solve his problems more quickly than Student B. The principal difference between these two kinds of information lies in the standard used as a reference. What I shall call criterion-referenced measures depend upon an absolute standard of quality, while what I term norm-referenced measures depend upon a relative standard (Berk, 1980, p. 3).

During the years immediately following the publication of Glaser's essay, interest in the newly defined arena of criterion-referenced testing remained at a minimum. Then in 1969, W. James Popham and

T. R. Husek, drawing on a very small body of relevant test development work by Hively (1968) and others in the intervening years, published the article, "Implications of Criterion-Referenced Measurement." This effort together with those of Popham and others in developing behavioral objectives helped give rise to full-blown interest in criterion-referenced testing. Regarding the impetus given criterion-referenced test development by the behavioral objectives movement, Popham (1978) concluded "that the more explicitly educators stated their objectives, the less suitable they perceived extant tests to be for assessing whether their objectives had been attained" (p. 15).

Since the early conceptual work of the 1960's, numerous definitions of criterion-referenced measurement and an abundance of synonymous terms have been generated. A sampling of these includes not only criterion-referenced test, but also domain-referenced test, objectives-referenced test, competency-based test, proficiency test, and mastery test (Berk, 1980, p. 4).

Among the definitions for criterion-referenced measurement which have multiplied since the late 1960's, Popham's (1978, p. 93) version summarizes many others. He states that "a criterion-referenced test is used to ascertain an individual's status with respect to a well-defined behavioral domain." This statement embodies the general agreement among various descriptions that a criterion-referenced test references an individual's score to a well-defined domain of behavior(s). Hence, the term domain-referenced test which some researchers (Lidz, 1981 and Roid & Haladyna, 1982) prefer over the more widely used term, criterion-referenced test. As for the term domain, the word is used alternately to refer to learning behaviors and to the universe of

possible test items for the particular content or behavior being tested (Lidz, 1981 and Roid & Haladyna, 1982). Accordingly, Lidz states:

> Although criterion-referenced measures do specify a criterion of mastery, they derive their name from...the meaning of criterion, which is a reference for domain. A criterion-referenced measure samples a larger domain of objectives that are considered to define the task to be learned. Because this domain of objectives is the criterion to which the test items are referenced, performance on the test is expected to accurately predict performance in the domain (p. 57).

Following a survey of definitions for the concept of criterion-referenced testing, the next task was to review what others have mapped out as constituting the process of test development. Extracting and combining from the stages given by Hambleton (Berk, 1980), Roid and Haladyna (1982), and Popham (1978), twelve (12) steps have been outlined.

1) Clarifying Instructional Intent - Roid and Haladyna (1982) in listing this as a distinct first step in criterion-referenced test development are referring to the traditional course content outline. The "Course Objectives" as they are termed for our Developmental Reading I course are an example of instructional intent in written form (see appendix). Of course, though we have labeled this enumeration of our instructional intent as "objectives," the statements do not actually merit the name behavioral objectives. They lack the specifics that Mager (1962) describes as requirements for performance objectives. These particulars include: 1) what the learner is doing when demonstrating achievement, 2) what the conditions are under which the learner is to perform, and 3) how well the learner is to perform.

2) Specifying Domain - This necessary phase may be accomplished by formulating properly constructed behavioral objectives or by engaging

BEST COPY AVAILABLE

in what is called task analysis or domain analysis (Lidz, 1981). In any case, clarification of instructional intent must take place. During this stage, those who favor objectives-referencing must write either a single behavioral objective or a set of learning objectives specifying the precise skill(s) and mastery level(s) expected of the learner. It is also necessary to specify the manner in which each objective is to be assessed, i.e., the test conditions. According to Popham (1978), this second stage of test development involves specifying a domain of behaviors

> which consists of a set of skills or dispositions that examinees display when called on to do so in a testing situation. For example, a behavioral domain in the field of mathematics might consist of an individual's ability to solve a certain class of simultaneous equation problems. A behavioral domain in the attitudinal realm might be the kinds of responses that an individual would make to forced-choice types of items in which one's preference for certain kinds of leisure activities would be reflected (p. 94).

Several challenges arise in the second stage although test constructors may not realize this until they arrive at later stages. One is the question of the generality of the behavioral domain being measured. Are the domains measured to be microscopic or macroscopic? One researcher articulates the question in this way:

> A behavioral domain can be defined so that it is quite complex in nature, consisting of high-level skills. On the other hand, a behavioral domain can be miniscule in nature - for example, testing the examinee's ability to spell correctly ten particular words (Popham, p. 94).

According to many experts, another problem that may later become obvious is that behavioral objectives as usually stated do not give sufficient guidance for test-item writing. What generally occurs is that given

the same behavioral objective different writers of test items will produce widely varying questions. Divergent test items in turn raise these questions: Do all variations in test-items produced serve as measures of the same things and do the various items equally reflect the objectives from which they were generated? Another crucial question is: What is the relative importance of an objective? Finally, one question concerning objectives supersedes all others: How or on what bases were the behavioral objectives derived? Roid and Haladyna (1982) and Lidz (1981) assert that these are among the most neglected areas of inquiry in criterion-referenced test development: How do we decide what is to be measured and determine the significance of what are chosen as objects of measurement?

3) Writing Test Specifications - Popham (1978) maintains that the usual 25 or so words constituting a particular objective do not sufficiently define the domain - what it is the test is to measure - nor do they tell test-item writers exactly what kinds of items are allowable in light of the objectives. This latter desired congruence between test-items and objectives is referred to as "derivative homogeneity" and when test items have this characteristic, it is possible to look at them and tell what domain is their rightful home. The scheme for writing specifications which is presented by Popham in Criterion-Referenced Measurement (1978) contains the following five components:

> 1) General description: A brief depiction, in general terms, of the behavior being assessed by the test.
>
> 2) Sample item: An illustrative item that reflects the test-item attributes to be delimited in the following two components.
>
> 3) Stimulus attributes: A series of statements that attempt to delimit the class of stimulus material that will be encountered by the examinee.

> 4) Response attributes: A series of statements that attempt either to a) delimit the classes of response options from which the student makes selected responses or b) explicate the standards by which an examinee's constructed responses will be judged.
>
> 5) Specification supplement: In certain cases it may be necessary to add an appendix or supplement to the preceding four components. This supplement typically provides a more detailed listing or explanation of of eligible content (pp. 121-122).

Other researchers such as Hambleton (Berk, 1980) list under test specifications such considerations as length of the proposed test, number of items per domain and whether objectives and items which represent the domain in the test will be chosen by random or stratified sampling (Lidz, 1981, p. 57).

4) Developing Test Items - In this step, a pool of test items is generated. These should follow standard rules for item writing and any specifications developed for a particular test. There are many books available which provide general rules for the writing of individual test items of various kinds. One of particular interest for those building criterion-referenced tests is a book by Roid and Haladyna entitled, A Technology for Test-Item Writing (1981). There are two chapters devoted to practical considerations for item writing.

5) Editing Test Items - At this stage, initial editing of test items is completed by those writing them. They review items to discover whether there may be any violations of the usual item-writing guidelines.

6) Determining Test Item Content Validity - Researchers identify two approaches for reviewing items. One is logical item review whose purpose is to insure congruence or consistency between items and the objectives from which they were drawn. For this type of review, content specialists may be called upon to inspect test items and to give their judgments. The other approach for determining content validity is

empirical review based on field testing. Berk in the chapter "Item Analysis" which he contributed to the book, Criterion-Referenced Testing: The State of the Art (for which he also served as editor), gives guidelines for empirical review procedures intended especially for practitioners. Though the methods suggested may be time-consuming, Berk says that they can be carried out without the aid of a computer if necessary.

7) Editing Test Items - Based on data obtained during the previous step, items are once again edited, questionable items being discarded if neither first aid nor major surgery improves their condition.

8) Assembling the Test - This phase of development includes such things as:

a) Determination of number of test items per domain specification.

b) Test item selection.

c) Preparation of directions and sample questions.

d) Layout and test booklet preparation.

e) Preparation of scoring keys.

f) Preparation of answer sheets (Hambleton, in Berk, 1980, p. 83).

9) Setting Performance Standards - In a chapter entitled "Test-Score Validity and Standard - Setting Methods" in Berk (1980), Hambleton reviews methods of standard setting. He includes both judgmental and empirical methods as well as combinations of these. Needless to say, the necessity of setting standards is one of the most involved problems in developing a criterion-referenced test. Popham also offers many practical strategies in the same volume.

10) Administering the Test - Field or pilot-testing is conducted during this stage. Groups of students must be chosen according to their

availability, the purpose of the test, and the methods chosen for gathering item statistics. At this time, technical assistance may be sought by developmental reading practioners. Liaisons may be formed with colleagues in other disciplines in order to insure that a sound instrument is developed.

11) Assessing Validity, Reliability - Data related to reliability and validity are gathered and analyzed. As a glance at the contents of the most recent issues of the Journal of Measurement will reveal, much is being done in the way of refining methods for assessing the validity and reliability of criterion-referenced tests, for which the means tailored to norm-referenced tests are not always appropriate. Practitioners would do well to consult the Hambleton chapter already referred to in which he deals with test score validity.

12) Preparing a User's Manual/Technical Manual - A user's manual and technical manual are prepared for distribution along with the test. Beyond this step there is a need to continue gathering technical data as the test is used in different situations and with different student populations.

Obviously, undertaking to develop a criterion-referenced test of reading for college freshmen requires an extensive commitment. A review of theoretical considerations makes clear that constructing a sound instrument is no casual venture. However, the survey of the literature summarized here reveals not only the formidable aspects of criterion-referenced test construction, but also that there is a wealth of work to date to guide and sustain the developmental reading practioner in such an effort.

REFERENCES

Berk, R.A., ed. Criterion-Referenced Measurement: The State of the Art. Baltimore: Johns Hopkins University Press, 1980.

Brown, J. I., Bennett, J. M., and Hanna, G., The Nelson-Denny Reading Test Froms E and F. Lombard, Illinois: Riverside, 1981.

Cummins, R. P., "Test Review: The Nelson-Denny Reading Test (Forms E and F)." Journal of Reading, 25, No. 1 (1981): pp. 54-59.

Curtis, M. E., and Glaser, R., "Reading Theory and the Assessment of Reading Achievement.: Journal of Educational Measurement, 20, No. 2 (1983): pp. 133-147.

Glaser, R., "Instructional Technology and the Measurement of Learning Outcomes: Some Questions." American Psychologist, 18 (1963): pp. 519-521.

Gronlund, N. E., Constructing Achievement Tests, 2nd ed. Englewood Cliffs, N.J.: Prentice-Hall, 1977.

Hoffman, B., The Tyranny of Testing. Westport, Conn.: Greenwood Press, 1962.

Lidz, C., Improving Assessment of Schoolchildren. San Francisco: Jossey-Bass, 1981.

Louisiana State University Developmental Reading Department, "Developmental Reading Courses with Exit Levels and Tests." Baton Rouge, 1982.

Mager, R.F., Preparing Instructional Objectives. 2nd ed. Blemont, Calif.: Fearon, 1975.

Owen, D., "1983: The Last Days of ETS." Harp[illegible]'s, May 1983: pp. 21-37.

Popham, W. J., Criterion-Referenced Measurement. Englewood Cliffs, N.J.: Prentice-Hall, 1978.

Popham, W. J., and Baker, E. L. Establishing Instructional Goals. Los Angeles, Calif.: Prentice-Hall, 1975.

Popham, W.J. and Husek, T.R. "Implications of Criterion-Referenced Measurement." Journal of Educational Measurement, 6 (1969): pp. 1-9.

Roid, G. H., and Haladyna, T. M. A Technology for Test-Item Writing. New York: Academic Press, 1982.

Appendix

COURSE OBJECTIVES

A. Vocabulary

That students develop a combination of methods for increasing their vocabularies independently by:

a. Learning and being able to recite the meanings of the set of combining forms presented in Gaining Word Power and to apply these in word analysis to derive meanings.

b. Using context clues to derive appropriate meanings of words.

c. Using dictionaries and glossaries to locate and verify word meanings and other items of information found in entries.

B. Comprehension

1. That students develop the skill of active previewing and surveying of textbook passages.

2. That students develop their ability to identify and recite the topics and main ideas of paragraphs and longer passages.

3. That students develop their ability to indentify and recite major and minor details which support main ideas in paragraphs and longer passages.

4. That students develop the skill of writing topic outlines of single and multiple paragraph passages for note-taking and study purposes.

5. That students develop the skill of composing summaries based on expository passages of several paragraphs.

6. That students develop the ability to draw correct inferences from reading.

7. That students develop the ability to read critically by detecting such things as an author's purpose, tone, use of words for connotative value, and use of opinion vs. fact.

8. That students develop their recognition of patterns of organization to guide comprehension by studying definition-examples, comparison-contrast, cause-effect, simple listings, and like patterns.

C. Directed Recreational Reading

A major goal of DEVP 90 is to encourage in students the attitude that reading can be a pleasurable, eye-opening, and mind-expanding experience. Therefore, students are asked to read at least two (2) novels, biographies, or autobiographies during the semester.